Biostatistics Definition

Biostatistics is the application of statistical techniques to scientific research in health-related fields, including medicine, biology, and public health, and the development of new tools to study these areas. Since the beginning of the twentieth century, the field of biostatistics has become an indispensable tool in improving health and reducing illness.

Biostatisticians play essential roles in designing studies, analyzing data and creating methods to attack research problems as diverse as:

- -the determination of major risk factors for heart disease, lung disease and cancer
- -the testing of new drugs to combat AIDS
- -the evaluation of potential environmental factors harmful to human health, such as tobacco smoke, asbestos or pollutants

Methods of Biostatistics

In biostatistics, for each of the specific situation, statistical methods are available for analysis and interpretation of the data. To select the appropriate statistical method, one needs to know the assumption and conditions of the statistical methods, so that proper statistical method can be selected for data analysis. Two main statistical methods are used in data analysis:

1. **Descriptive statistics**- which summarizes data using indexes such as mean and median.
2. **Inferential statistics**- which draw conclusions from data using statistical tests such as student's *t*-test.

Principles of Biostatistics

Module Aims- This core module aims to provide students with the necessary knowledge and biostatistical skills to be able to interpret and conduct basic statistical analyses of population health data. Students may choose to take this, or Statistics for HDS.

Module Learning Outcomes- By the end of the module, students should be able to:

1. Understand sampling variation in the context of population health studies
2. Use R to manipulate data and to apply and interpret the output of commonly used statistical procedures
3. Select, perform and interpret appropriate descriptive analyses of population health data
4. Select, apply and interpret common regression models for the statistical analysis of population health data
5. Perform standard sample size and power calculations

Pre-requisites- Fluent numeracy and a good understanding of elementary algebra (e.g. rearranging equations, graphical interpretation of a linear equation in two variables, simultaneous linear equations when the solution is unique, quadratic equations), logarithms (and performing operations on logarithmic scale), summation notation (∑), and probability (including performing simple probability operations). Familiarity with scientific notation, and with performing automated calculations (e.g. in excel, R, or equivalent).

Teaching Strategy- The module will be delivered using a combination of lectures, class discussion, small-group exercises, and computer practical. Some reading may be required prior to some sessions.

Assessment- A timed open book assessment approximately 1 week after the end of the module, consisting of a data analysis task and associated report.

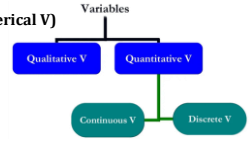
Variable:

Characteristic that can take on different values for different persons, places or things. Exp: Age, Gender, Blood pressure.

Variable generally two types_

I.Quantitative Variable (Numerical V)

Measurements made on quantitative variables convey information regarding amount. Quantitative Variable two sub types_



a) Discrete V (Countable): Is characterized by gaps or interruptions in the values that it can assume.

b) Continuous V (Measurable): can assume any value within a specified relevant interval of values.

II. Qualitative (Nominal) Variable:

Some characteristics are not capable of being measured in the sense, and can be categorized only.

Measurements in biostatistics

Measurement is the assigning of numbers and codes according to prior-set rules (Stevens, 1946).

Three main types of measurements:

1. **Categorical (nominal) -** Classify observations into named categories. Exp- HIV status, SEX (male or female) etc.
2. **Ordinal-** Categories that can be put in rank order. Exp- STAGE OF CANCER classified as stage I, stage II, stage III, stage IV
3. **Quantitative (scale) -** Numerical values with equal spacing between numerical values (like number line). Exp- AGE (years), SERUM CHOLESTEROL (mg/dL), T4 cell count (per dL)

Functions of Statistics

1. **Statistics simplifies complexity:** Statistic consists of aggregate of numerical facts. Huge facts and figures are difficult to remember. The complex mass of figures can be made simple and understandable with the help of statistical methods
2. **Statistics presents fact in a definite form**: One of the important functions of statistics is to present the general statements in a precise and definite form.
3. **Statistics facilities comparison:** The science of statistics does not mean only counting but also comparison. Unless the figures are compared with other figures with the same kind, they are meaningless
4. **To help in formulation of policies**: Statistics helps in formulating the policies in different fields mainly in economics, business etc.
5. **Statistics helps in forecasting**: While preparing suitable policies and plans, it is necessary to have the knowledge of future tendency.
6. **Statistics helps in formulating and testing hypothesis:** Statistical methods are helpful not only in estimating the present forecasting the future but also helpful in formulating and testing the hypothesis for the development of new theories.

Limitations of Statistics

Besides the importance of statistics in every field of life, it has some limitations. The following are the main limitations of statistics are:

1. **Statistics does not deal with individuals:** A part of the definition of statistics is that it must be the aggregates of facts.
2. **Statistics does not study qualitative phenomena**: The science of statistics studies only the quantitative aspect of the problem.
3. **Statistical laws are not exact:** 100% accuracy is rare in statistical work because statistical laws are true only on the average.
4. **Statistics is only a means:** Statistical methods provide only a method of studying problem. There are other methods also. These methods should be used to supplement the conclusions derived with the help of statistics.
5. **Statistics is liable to be misused:** The most important limitation of statistics is that it must be handled by experts. Statistical methods are the most dangerous tools in the hands of inexpert.

Uses of Biostatistics

1. **Collection of data:** For any statistical investigation, the first tool to be used is the collection of data. The data may be primary or secondary. The collection of primary or secondary data depends upon the nature, object, scope of the enquiry, financial resource, time factor and the degree of accuracy. The result of the analysis and its interpretation totally depend upon the data collected. So, the data must be collected carefully.
2. **Organization:** After completing the process of collecting the data, the second tool to be used is the method of organization. Organization of the data depends upon the source from which the data are obtained. If the data are obtained from the published source, it will generally be in the organized form.
3. **Presentation:** After collection and organizing the data, the next tool to be used is to present them systematically so that they can be presented in various forms such as table form, diagrammatic form and graphical form. With the help of this tool, comparison between two can be made easily.
4. **Analysis:** After collection, organization and presentation of the data, the important step to be used is the analysis of data. Various statistical tools such as averages, dispersion, correlation, test of significance, index number, time series etc. can be used to analyze the data. Statistical tools or appropriate technique depends upon the nature of the data and the purpose of the inquiry.

Data Collection

In Statistics, data collection is a process of gathering information from all the relevant sources to find a solution to the research problem. It helps to evaluate the outcome of the problem. The data collection methods allow a person to conclude an answer to the relevant question. Most of the organizations use data collection methods to make assumptions about future probabilities and trends. Once the data is collected, it is necessary to undergo the data organization process.

The main sources of the data collections methods are "Data". Data can be classified into two types, namely primary data and secondary data.

1. **Primary Data Collection Methods:**

Primary data or raw data is a type of information that is obtained directly from the first-hand source through experiments, surveys or observations. The primary data collection method is further classified into two types. They are

a) **Quantitative Data Collection Methods** - It is based on mathematical calculations using various formats like close-ended questions, correlation and regression methods, mean, median or mode measures.

b) **Qualitative Data Collection Methods** - It does not involve any mathematical calculations. This method is closely associated with elements that are not quantifiable. This qualitative data collection method includes interviews, questionnaires, observations, case studies, etc.

2. **Secondary Data Collection Methods :**

Secondary data is data collected by someone other than the actual user. It means that the information is already available, and someone analyses it. The secondary data includes magazines, newspapers, books, journals, etc.

Merits and Demerits of Data Collation

Merits	Demerits
1. **Economical**: It is economical, because we have not to collect all data. Instead of getting data from 5000 farmers, we get it from 50-100 only. 2. **Less Time Consuming:** As no of units is only a fraction of the total universe, time consumed is also a fraction of total time. 3. **Reliable**: If sample is taken judiciously, the results are very reliable and accurate. 4. **Organisational Convenience:** As samples are taken and the number of units is smaller, the better (Trained) enumerators can be employed by the organisation. **5. More Scientific**: According to Prof R.A. Fisher, "The sample technique has four important advantages over census technique of data collection.	**1. Absence of Being Representative**: Methods, such as purposive sampling may not provide a sample that is representative. **2. Wrong Conclusion**: If the sample is not representative, the results will not be correct. These will lead to the wrong conclusions. **3. Small Universe:** Sometimes universe is so small that proper samples cannot be taken not of it. Number of units are so less. 4. **Specialised Knowledge**: It is a scientific method. Therefore, to get a good and representative sample, one should have special knowledge to get good sample and to perform proper analysis so that reliable result may be achieved. 5. **Inherent defects:** The results which are achieved though the analysis of sampling data may not be accurate as this method have inherent defects.

PRESENTATION OF STATISTICAL DATA:

Statistical data can be presented in three different ways_

(a) Textual presentation: This is a descriptive form. The following is an example of such a presentation of data about deaths from industrial diseases in Great Britain in 1935–39 and 1940-44.

(b) Tabular presentation, or, Tabulation: Tabulation may be defined as the systematic presentation of numerical data in rows or/and columns according to certain characteristics. It expresses the data in concise and attractive form which can be easily understood and used to compare numerical figures. Before drafting a table, you should be sure what you want to show and who will be the reader.

(c) Graphical presentation: Quantitative data may also be presented graphically by using bar charts, pie diagrams, pictographs, line diagrams, etc.

Classification of Data

The collected data, also known as raw data or ungrouped data are always in an un organised form and need to be organised and presented in meaningful and readily comprehensible form in order to facilitate further statistical analysis. It is, therefore, essential for an investigator to condense a mass of data into more and more comprehensible and assimilable form. The process of grouping into different classes or sub classes according to some characteristics is known as classification, tabulation is concerned with the systematic arrangement and presentation of classified data. Thus classification is the first step in tabulation.

Objects of Classification:

The following are main objectives of classifying the data:

1. It condenses the mass of data in an easily assimilable form.
2. It eliminates unnecessary details.
3. It facilitates comparison and highlights the significant aspect of data.
4. It enables one to get a mental picture of the information and helps in drawing inferences.
5. It helps in the statistical treatment of the information collected.

Types of classification:

Statistical data are classified in respect of their characteristics. Broadly there are four basic types of classification namely

1. Chronological classification – In chronological classification the collected data are arranged according to the order of time expressed in years, months, weeks, etc..
2. Geographical classification - In this type of classification the data are classified according to geographical region or place.
3. Qualitative classification - In this type of classification data are classified on the basis of same attributes or quality like sex, literacy, religion, employment etc., Such attributes cannot be measured along with a scale.
4. Quantitative classification- Quantitative classification refers to the classification of data according to some characteristics that can be measured such as height, weight, etc.

Tabulation:

Tabulation is the process of summarizing classified or grouped data in the form of a table so that it is easily understood and an investigator is quickly able to locate the desired information. A table is a systematic arrangement of classified data in columns and rows. Thus, a statistical table makes it possible for the investigator to present a huge mass of data in a detailed and orderly form. It facilitates comparison and often reveals certain patterns in data which are otherwise not obvious. Classification and 'Tabulation', as a matter of fact, are not two distinct processes. Actually they go together. Before tabulation data are classified and then displayed under different columns and rows of table_

Immunization status	Number	Percent
Not immunized	75	35.7
Partially immunized	57	27.1
Fully immunized	78	37.2
Total	210	100.0

Sampling methods

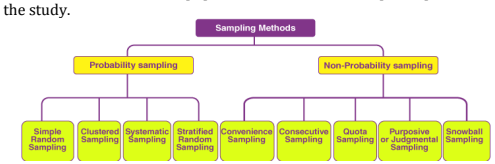
In Statistics, the sampling method or sampling technique is the process of studying the population by gathering information and analyzing that data. It is the basis of the data where the sample space is enormous.

Types of Sampling Method

In Statistics, there are different sampling techniques available to get relevant results from the population. The two different types of sampling methods are_

1. Probability Sampling- The probability sampling method utilizes some form of random selection. In this method, all the eligible individuals have a chance of selecting the sample from the whole sample space. This method is more time consuming and expensive than the non-probability sampling method.

2. Non-probability Sampling- The non-probability sampling method is a technique in which the researcher selects the sample based on subjective judgment rather than the random selection. In this method, not all the members of the population have a chance to participate in the study.



Unit-3

Mean median and mode are the three measures of central tendency. Mean is the most commonly used measure of central tendency. It actually represents the average of the given collection of data. It is applicable for both continuous and discrete data. It is equal to the sum of all the values in the collection of data divided by the total number of values. **Mean is calculated for ungrouped** data using the formula as
Mean =

In the case of **grouped** data, the mean is calculated using three methods such as: Direct method, Assumed mean method and Step deviation method.

Merits of mean:
1) Arithmetic mean rigidly defined by Algebraic Formula.
2) It is easy to calculate and simple to understand.
3) It is based on all observations of the given data.
4) It is capable of being treated mathematically hence it is widely used in statistical analysis.
5) Arithmetic mean can be computed even if the derailed distribution is not known but some of the observation and number of the observation are known.
Demerits of mean :
1) It can neither be determined by inspection or by graphical location.
2) Arithmetic mean can not be computed for qualitative data like data on intelligence honesty and smoking habit etc.
3) It is too much affected by extreme observations and hence it is not adequately represent data consisting of some extreme point.
4) Arithmetic mean can not be computed when class intervals have open ends.
5) If any one of the data is missing then mean can not be calculated.

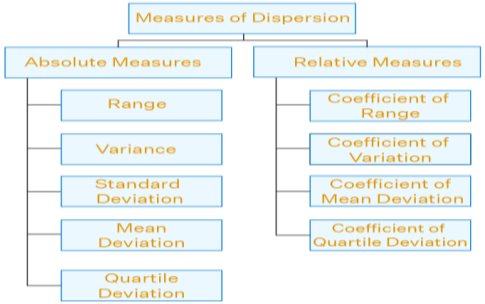
Median is the middle entry in the sorted sequence. For example, the median of 1, 3, 4, 10, 17 is 4. If there is an even number of values, then just take the average of those two. For example, the median of 1, 3, 8, 10, 21, 25 is the average of 8 and 10, that is, 9.
For ungrouped data, the median can be calculated using the formulas given below:
Median = L+ [(N/2 – cf)/f] × h
Here
l = lesser limit belonging to the median class
c = cumulative frequency value of the class before the median class
f = frequency possessed by the median class
h = size of the class

Merits or Uses of Median:
1. Median is rigidly defined as in the case of Mean.
2. Even if the value of extreme item is much different from other values, it is not much affected by these values.
3. It can also be used for the Quantities; those can't give A.M; as is in case of intelligence etc.
4. It can be located graphically.
Demerits or Limitations of Median:
1. Even if the value of extreme items is too large, it does not affect too much, but due to this reason, sometimes median does not remain the representative of the series.
2. It is affected much more by fluctuations of sampling than A.M.
3. Median cannot be used for further algebraic treatment. Unlike mean we can neither find total of terms as in case of A.M. nor median of some groups when combined.
4. If the number of series is even, we can only make its estimate; as the A.M. of two middle terms is taken as Median.

Mode is the number that appears most often. For example, given 1, 1, 2, 2, 2, 3, 3, the mode is 2.
Mode = 3 Median – 2 Mean

Merits or Uses of Mode:
1. Mode is the term that occur most in the series hence it is not an isolated value like Median nor it is value like mean that may not be there in the series.
2. It is not affected by extreme values hence is a good representative of the series.
3. It can be found graphically also.
4. It can also be used in case of Quantitative phenomenon.
Demerits or Limitations of Mode:
1. Mode cannot be determined if the series is bimodal or multimodal.
2. Mode is based only on concentrated values; other values are not taken into account in-spite of their big difference with the mode.
3. Mode is most affected by fluctuation of sampling.
4. Mode is not so rigidly defined. Solving the problem by different methods we won't get the same results as in case of mean.

Measures of dispersion
Measures of dispersion help to describe the variability in data. Dispersion is a statistical term that can be used to describe the extent to which data is scattered. Thus, measures of dispersion are certain types of measures that are used to quantify the dispersion of data. Measures of dispersion can be defined as positive real numbers that measure how homogeneous or heterogeneous the given data is. The value of a measure of dispersion will be 0 if the data points in a data set are the same. However, as the variability of the data increases the value of the measures of dispersion also increases.
Types of Measures of Dispersion
The measures of dispersion can be classified into two broad categories. These are absolute measures of dispersion and relative measures of dispersion. Range, variance, standard deviation and mean deviation fall under the category of absolute measures of deviation.



The range in statistics for a given data set is the difference between the highest and lowest values.
Thus, the range could also be defined as the difference between the highest observation and lowest observation. The obtained result is called the range of observation. The range in statistics represents the spread of observations.
Range Formula
The formula of the range in statistics can simply be given by the difference between highest and lowest value.
Range = Highest Value – Lowest Value
Find Range in Statistics
To find the range in statistics, we need to arrange the given values or set of data or set of observations in ascending order. That means, firstly write the observations from lowest to highest value. Now, we need to use the formula to find the range of observations.

Merits or Uses:
1. It is easiest to calculate and simplest to understand even for a beginner.
2. It is one of those measures which are rigidity defined.
3. It gives us the total picture of the problem even with a single glance.
4. It is used to check the quality of a product for quality control. Range plays an important role in preparing R- charts, thus quality is maintained.
5. The idea about the price of Gold and Shares is also made taking care of the range in which prices have moved for the past some periods.

Demerits or Limitations or Drawbacks:
1. Range is not based on all the terms. Only extreme items reflect its size. Hence range cannot be completely representative of the data as all other middle values are ignored.
2. Due to above reason range is not a reliable measure of dispersion.
3. Range does not change even the least even if all other, in between, terms and variables are changed.
4. It does not tell us anything about the variability of other data.

Standard Deviation
The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean.

The Formula for Standard Deviation

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

where:
x_i = Value of the i^{th} point in the data set
$\bar{x}$ = The mean value of the data set
n = The number of data points in the data

Merits
1. It is rigidly defined and free from any ambiguity.
2. Its calculation is based on all the observations of a series and it cannot be correctly calculated ignoring any item of a series.
3. It strictly follows the algebraic principles, and it never ignores the + and – signs like the mean deviation.
4. It is capable of further algebraic treatment as it has a lot of algebraic properties.

Demerits
1. It is not understood by a common man.
2. Its calculation is difficult as it involves many mathematical models and processes.
3. It is affected very much by the extreme values of a series in as much as the squares of deviations of big items proportionately bigger than the squares of the smaller items.
4. It cannot be used for comparing the dispersion of two or more series given in different units.

Mean Deviation Definition
The mean deviation is defined as a statistical measure that is used to calculate the average deviation from the mean value of the given data set. The mean deviation of the data values can be easily calculated using the below procedure.
Step 1: Find the mean value for the given data values
Step 2: Now, subtract the mean value from each of the data values given (Note: Ignore the minus symbol)
Step 3: Now, find the mean of those values obtained in step 2.

Mean Deviation Formula
The formula to calculate the mean deviation for the given data set is given below.
Mean Deviation = [Σ |X – μ|] /N
Here,
Σ= represents the addition of values
X= represents each value in the data set
μ= represents the mean of the data set
N= represents the number of data values
Merits
1. It is simple to understand.
2. It is easy to calculate.
3. It is based on all the observations of a series.
4. It shown the dispersion, or scatter of the various items of a series from its central value.
5. It is not very much affected by the values of extreme items of a series.

Demerits
1. It is not rigidly defined in the sense that it is computed from any central value viz. Mean, Median, Mode etc. and thereby it can produce different results.
2. It violates the algebraic principle by ignoring the + and – signs while calculating the deviations of the different items from the central value of a series.
3. It is not capable of further algebraic treatment.
4. It is affected much by the fluctuations in sampling.
5. It is not suitable for sociological study.

Quartile Deviation
The Quartile Deviation can be defined mathematically as half of the difference between the upper and lower quartile. Here, quartile deviation can be represented as QD; Q3 denotes the upper quartile and Q1 indicates the lower quartile.
Quartile Deviation is also known as the Semi Interquartile range.
Quartile Deviation Formula
Suppose Q1 is the lower quartile, Q2 is the median, and Q3 is the upper quartile for the given data set, then its quartile deviation can be calculated using the following formula.
Q.D. = (Q3 – Q1)/2
In the next section, you will learn how to calculate these quartiles for both ungrouped and grouped data separately.
Merits of Quartile Deviation:
1. It can be easily calculated and simply understood.
2. It does not involve much mathematical difficulties.
3. As it takes middle 50% terms hence it is a measure better than Range and Percentile Range.
4. It is not affected by extreme terms as 25% of upper and 25% of lower terms are left out.
5. In case we are to deal with the center half of a series this is the best measure to use.

Demerits or Limitation Quartile Deviation:
1. As Q1 and Q3 are both positional measures hence are not capable of further algebraic treatment.
2. Calculation is much more, but the result obtained is not of much importance.
3. It is too much affected by fluctuations of samples.
4. If the values are irregular, then result is affected badly.
5. We can't call it a measure of dispersion as it does not show the scatterness around any average.

Co- efficient of variations

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.
Below is the formula for how to calculate the coefficient of variation:
CV = μ/σ x 100
Where
σ=standard deviation, μ=mean
Please note that if the expected return in the denominator of the coefficient of variation formula is negative or zero, the result could be misleading.

Example of Coefficient of Variation
Fred wants to find a new investment for his portfolio. He is looking for a safe investment that provides stable returns. He considers the following options for investment:
Stocks: Fred was offered stock of ABC Corp. It is a mature company with strong operational and financial performance. The volatility of the stock is 10%, and the expected return is 14%.
ETFs: Another option is an Exchange-Traded Fund (ETF) which tracks the performance of the S&P 500 index. The ETF offers an expected return of 13% with a volatility of 7%.
Bonds: Bonds with excellent credit ratings offer an expected return of 3% with 2% volatility.

Uni-4

Correlation is a statistical calculation that indicates that two variables are parallelly related (which means that the variables change together at a constant rate). It is a simple and popularly used tool for defining relationships without delivering a statement concerning the cause and effect.
In simple words, correlation is a statistical calculation that estimates the point at which the two variables shift in relation to each other.

Types of Correlation:

Depending upon the direction and proportion of changes in the variables and the number of data series, the correlation may be of the following types:

1. Positive and Negative Correlations:

When the changes take place in same direction in two variables or data series, the correlation between them is said to be positive and direct. For example, if increase in one variable causes increase in the other variable or a decrease in one variable causes decrease in the other variable, the two variables show positive correlation.

2. Linear and Non-Linear Correlations:

When the proportion of changes in the two variables or data series is fixed, i.e., the dispersions in two series are equal, the correlation is said to be linear. If two such variables are plotted on O-X and O-y axes of graph paper, the changes in two will result in straight line graph.

3. Simple, Multiple and Partial Correlation:

The correlation existing between two variables or data series is said to be simple correlation. Of the two series, one which causes change in the other is called independent or subject series and the other which is affected is called dependent series.

Methods of Determining Correlation:

The following methods are generally used to determine simple correlation:

a. Graphic Method:

When the values of dependent series are plotted on O –X axis and independent series are plotted on O-Y axis of graph paper, a linear or non-linear graph will be obtained which will simply indicate the direction of correlation and not the numerical magnitude.

b. Scatter Diagram or Dotogram Method:

This method is more or less similar to graphic method. In this method, the values of independent data series are plotted on O – X axis and those of dependent series on 0-Y axis and then the pairs of values are plotted on the graph paper.

C. Karl Pearson's Coefficient of Correlation Method:

This is the best mathematical method of determining the correlation. Coefficient of correlation (r) is obtained by dividing the product of values of covariance of the two series by the product of their standard deviations.

$$r = \frac{Cov.(X.Y)}{\sigma_X . \sigma_Y}$$

Where σX and σY are the standard deviation of variables of data series, X and Y

d. Spearman's Ranking Method:

Professor Charls Spearman worked out a method for determining correlation in which the values of all data of a series are assigned ranks in decreasing or increasing (ascending) order. In this ranking process, the highest value is given rank 1 and the next higher value is given rank 2 and so on. In some series the values of two or more data are similar.
In that case, the mean of the ranks will be equally shared by those data, as for example in one series there are two observations; one at S. No. 3 and the other at S. No. 10 of 67 each. In ranking process 67 at S. No. 3 and 67 at S. No. 10 instead of being ranked 6 and 7 respectively are ranked at 6.5 (mean of rank 6 and rank 7).
In the same way if there are three or more data in a series as have got same value, all those data will share the rank which will be the mean of their ranks. The number or frequency of the data with similar value is indicated by m.
In the next step, the difference between the ranks (D) of respective data of the two series arc obtained (D = R1-R2) which may be positive or negative figures. Then after, the values of D2 and sum of D2 (= ∑D2) are determined.
For two such series as are taking in data with similar values, the following formula is used to determine the coefficient of correlation by ranking method (Symbolized by Rho = ρ):

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).
Regression helps investment and financial managers to value assets and understand the relationships between variables, such as commodity prices and the stocks of businesses dealing in those commodities.
The general form of each type of regression is:
Simple linear regression: Y = a + bX + u
Multiple linear regression: Y = a + b1X1 + b2X2 + b3X3 + ... + btXt + u
Where:

Y = the variable that you are trying to predict
X = the variable that you are using to predict Y
a = the intercept.
b = the slope.
u = the regression residual.

Simple linear regression is used to estimate the relationship between two quantitative variables. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

Fitting prediction

A fitted value is a statistical model's prediction of the mean response value when you input the values of the predictors, factor levels, or components into the model. Suppose you have the following regression equation: y = 3X + 5. If you enter a value of 5 for the predictor, the fitted value is 20. Fitted values are also called predicted values.
The general procedure for using regression to make good predictions is the following:

1. Research the subject-area so you can build on the work of others.
2. Collect data for the relevant variables.
3. Specify and assess your regression model.
4. If you have a model that adequately fits the data, use it to make predictions.

Difference Between Correlation And Regression

Correlation	Regression
'Correlation' as the name says it determines the interconnection or a co-relationship between the variables.	'Regression' explains how an independent variable is numerically associated with the dependent variable.
In Correlation, both the independent and dependent values have no difference.	However, in Regression, both the dependent and independent variable are different.
The primary objective of Correlation is, to find out a quantitative/numerical value expressing the association between the values.	When it comes to regression, its primary intent is, to reckon the values of a haphazard variable based on the values of the fixed variable.
Correlation stipulates the degree to which both of the variables can move together.	However, regression specifies the effect of the change in the unit, in the known variable(p) on the evaluated variable (q).
Correlation helps to constitute the connection between the two variables.	Regression helps in estimating a variable's value based on another given value.

Similarities between correlation and regression

In addition to differences, there are some key similarities between correlation and regression that can help you to better understand your data.
1. Both work to quantify the direction and strength of the relationship between two numeric variables.
2. Any time the correlation is negative, the regression slope (line within the graph) will also be negative.
3. Any time the correlation is positive, the regression slope (line within the graph) will be positive.

Unit-5

Hypothesis

A hypothesis is an assumption that is made based on some evidence. This is the initial point of any investigation that translates the research questions into predictions. It includes components like variables, population and the relation between the variables. A research hypothesis is a hypothesis that is used to test the relationship between two or more variables.

Sources of Hypothesis

Following are the sources of hypothesis:

1. The resemblance between the phenomenon.
2. Observations from past studies, present-day experiences and from the competitors.
3. Scientific theories.
4. General patterns that influence the thinking process of people.

Characteristics of Hypothesis

Following are the characteristics of the hypothesis_

- The hypothesis should be clear and precise to consider it to be reliable.
- If the hypothesis is a relational hypothesis, then it should be stating the relationship between variables.
- The hypothesis must be specific and should have scope for conducting more tests.
- The way of explanation of the hypothesis must be very simple and it should also be understood that the simplicity of the hypothesis is not related to its significance.

Types of Hypothesis

There are six forms of hypothesis and they are:

1. **Simple Hypothesis-** It shows a relationship between one dependent variable and a single independent variable. For example – If you eat more vegetables, you will lose weight faster. Here, eating more vegetables is an independent variable, while losing weight is the dependent variable.
2. **Complex Hypothesis -** It shows the relationship between two or more dependent variables and two or more independent variables. Eating more vegetables and fruits leads to weight loss, glowing skin, and reduces the risk of many diseases such as heart disease.
3. **Directional Hypothesis** - It shows how a researcher is intellectual and committed to a particular outcome. The relationship between the variables can also predict its nature. For example- children aged four years eating proper food over a five-year period are having higher IQ levels than children not having a proper meal. This shows the effect and direction of the effect.
4. **Non-directional Hypothesis**- It is used when there is no theory involved. It is a statement that a relationship exists between two variables, without predicting the exact nature (direction) of the relationship.
5. **Null Hypothesis-** It provides a statement which is contrary to the hypothesis. It's a negative statement, and there is no relationship between independent and dependent variables. The symbol is denoted by "HO".
6. **Associative and Causal Hypothesis** - Associative hypothesis occurs when there is a change in one variable resulting in a change in the other variable. Whereas, the causal hypothesis proposes a cause and effect interaction between two or more variables.

Functions of Hypothesis

Following are the functions performed by the hypothesis:

- Hypothesis helps in making an observation and experiments possible.
- It becomes the start point for the investigation.
- Hypothesis helps in verifying the observations.
- It helps in directing the inquiries in the right direction.

Examples of Hypothesis

Following are the examples of hypotheses based on their types:

1. Consumption of sugary drinks every day leads to obesity is an example of a simple hypothesis.
2. All lilies have the same number of petals is an example of a null hypothesis.
3. If a person gets 7 hours of sleep, then he will feel less fatigue than if he sleeps less. It is an example of a directional hypothesis.

Simple Hypothesis

A simple hypothesis is a hypothesis that there exists a relationship between two variables. One is called a dependent variable, and the other is called an independent variable.
Simple hypotheses are ones which give probabilities to potential observations. The contrast here is with *complex* hypotheses, also known as *models*, which are sets of simple hypotheses such that knowing that some member of the set is true (but not which) is insufficient to specify probabilities of data points.

Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.
There are 5 main steps in hypothesis testing:

1. State your research hypothesis as a null hypothesis and alternate hypothesis (Ho) and (Ha or H1).
2. Collect data in a way designed to test the hypothesis.
3. Perform an appropriate statistical test.
4. Decide whether to reject or fail to reject your null hypothesis.
5. Present the findings in your results and discussion section.

Though the specific details might vary, the procedure you will use when testing a hypothesis will always follow some version of these steps.

Student's t-test, in statistics, a method of testing hypotheses about the mean of a small sample drawn from a normally distributed population when the population standard deviation is unknown.
In 1908 William Sealy Gosset, an Englishman publishing under the pseudonym Student developed the *t*-test and *t* distribution.
The *t* distribution is a family of curves in which the number of degrees of freedom specifies a particular curve.
T-test uses means and standard deviations of two samples to make a comparison. The formula for T-test is given below:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_\Delta}$$

where

$$s_\Delta = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where,
$\bar{x}$ = Mean of first set of values
$\bar{x}_2$ = Mean of second set of values
S_1 = Standard deviation of first set of values
S_2 = Standard deviation of second set of values
n_1 = Total number of values in first set
n_2 = Total number of values in second set.

The formula for standard deviation is given by:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Where,
x = Values given
$\bar{x}$ = Mean
n = Total number of values.

Chi-Square Test

A ***chi-squared test*** (symbolically represented as $\chi 2$) is basically a data analysis on the basis of observations of a random set of variables. Usually, it is a comparison of two statistical data sets. This test was introduced by ***Karl Pearson*** in 1900 for categorical data analysis and distribution. So it was mentioned as ***Pearson's chi-squared test***.
The chi-square test is used to estimate how likely the observations that are made would be, by considering the assumption of the null hypothesis as true.

Formula

The chi-squared test is done to check if there is any difference between the observed value and expected value. The formula for chi-square can be written as;

$$\chi 2 = \sum (Oi - Ei)2/Ei$$

Where, **Oi** is the observed value and **Ei** is the expected value.

Chi-Square Test of Independence

The chi-square test of independence also known as the chi-square test of association which is used to determine the association between the categorical variables. It is considered as a non-parametric test. It is mostly used to test statistical independence.
The chi-square test of independence is not appropriate when the categorical variables represent the pre-test and post-test observations. For this test, the data must meet the following requirements:

- Two categorical variables
- Relatively large sample size
- Categories of variables (two or more)
- Independence of observations